DOCUMENT RESUME

ED 400 274 TM 025 526

AUTHOR Kim, Seock-Ho; Cohen, Allan S.
TITLE A Comparison of Linking and Concurrent Calibration under Item Response Theory.
PUB DATE Apr 96
NOTE 54p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Adaptive Testing; Comparative Analysis; Computer Assisted Testing; Difficulty Level; Equated Scores; *Estimation (Mathematics); Item Bias; *Item Response Theory; *Testing Problems; *Test Items
IDENTIFIERS *Calibration; Item Characteristic Function; Item Parameters; *Linking Metrics; Marginal Maximum Likelihood Statistics

ABSTRACT

Applications of item response theory to practical testing problems including equating, differential item functioning, and computerized adaptive testing, require that item parameter estimates be placed onto a common metric. In this study, three methods for developing a common metric under item response theory are compared: (1) linking separate calibrations using equating coefficients from the characteristic curve method; (2) concurrent calibration via marginal maximum "a posteriori"; estimation; and (3) concurrent calibration via marginal maximum likelihood estimation. Linking using the characteristic curve method yielded smaller root mean square differences for both item discrimination and difficulty parameters for smaller numbers of common items. For the larger number of common items, the three methods yielded essentially the same results. (Contains 4 figures, 18 tables, and 23 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 274

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Seock-Ho Kim
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

A Comparison of Linking and Concurrent Calibration Under Item Response Theory

Seock-Ho Kim
The University of Georgia
Allan S. Cohen
University of Wisconsin–Madison

Running Head: LINKING AND CONCURRENT CALIBRATION

Paper presented at the annual meeting of the American Educational Research Association at New York, April, 1996.

TM 025526

BEST COPY AVAILABLE

A Comparison of Linking and Concurrent Calibration Under Item Response Theory

Abstract

Applications of item response theory to practical testing problems including equating, differential item functioning, and computerized adaptive testing, require item parameter estimates be placed onto a common metric. In this study, we compared three methods for developing a common metric under item response theory: (1) linking separate calibrations using equating coefficients from the characteristic curve method, (2) concurrent calibration via marginal maximum a posteriori estimation, and (3) concurrent calibration via marginal maximum likelihood estimation. Linking using the characteristic curve method yielded smaller root mean square differences for both item discrimination and difficulty parameters for smaller numbers of common items. For the larger numbers of common items, the three methods yielded essentially the same results.

Key words: BILOG, concurrent calibration, equating, linking, MULTILOG.

Introduction

Studies of horizontal and vertical equating and studies of differential item functioning under item response theory (IRT) require that item parameters from two or more data sets be expressed on a common metric. For purposes of this paper, we shall refer to linking as developing a common metric in IRT by transforming a set of item parameter estimates from one metric onto another, base metric. A common metric in IRT can also be constructed by simultaneously calibrating a combined data set. In spite of the fact that the metric of the θ scale is important under IRT, however, results from linking and concurrent calibration and the issue of the identification problem in these contexts have not been studied. In this study, therefore, we compare linking and concurrent calibration methods used for developing a common ability metric.

The purpose of equating is to convert test scores obtained from one test to the metric of another test. In horizontal equating, the tests to be equated are at the same level of difficulty and the ability distributions of examinees are comparable. Horizontal equating is required where multiple forms of a test are needed. In vertical equating, the tests to be equated are at the different levels of difficulty and the ability distributions of examinees are not comparable. Vertical equating is required so that a single scale can be used to make comparisons of abilities of examinees at different levels (e.g., different grades). Under IRT, equating may not be necessary, if item parameters from two tests are on the same metric. Hence, in IRT the task of equating is reduced to developing a common metric.

Both equating of test scores from various tests and linking of item parameters can be carried out under several different designs (Vale, 1986).

The focus in this paper is on the anchor test design in which two tests contain a set of common items and the tests are administered to two groups of examinees either with comparable or different ability levels.

When separate calibrations are used for dichotomously scored IRT models, three classes of linking methods are available for obtaining the linking or equating coefficients, A and B: characteristic curve methods (Divgi, 1980; Haebara, 1980; Stocking & Lord, 1983), the minimum chi-square method (Divgi, 1985), and mean and sigma methods (Linn, Levine, Hasting, & Wardrop, 1981; Loyd & Hoover, 1980; Marco, 1977; Stocking & Lord, 1983). The transformation coefficients, A and B, are obtained from the item parameter estimates of the common items on the two tests. In general, if there are two sets of item parameter estimates, one set from the base group and the other from the target group, the task is to place item and ability estimates of the target group onto the metric of the base group. Item parameter estimates from the target group, including those for the common items, are placed onto the metric of the base group via these coefficients. After the metric transformation and in order to achieve symmetry of transformation, the item parameter estimates from the base group and the transformed item parameter estimates from the target group for the common items are averaged to obtain the final estimates (Hambleton & Swaminathan, 1985).

Concurrent calibration is presently possible using two different estimation procedures, joint maximum likelihood estimation (JMLE) and marginal maximum likelihood estimation (MMLE). Bayes estimation can also be employed in either JMLE or MMLE contexts. JMLE is implemented in a number of computer programs including LOGIST (Wingersky, Barton, & Lord, 1982) and MICROSCALE (Mediax Interactive Technologies, 1985).

Concurrent calibration involves estimating item and ability parameters with a single run, combining data from both or several groups and treating items not taken by a particular group as not reached or missing (Lord, 1980). A variation of this is also possible in which the parameter estimates of the common items from the base group are set to be fixed and the remaining item parameters are estimated using data from the target group.

In the context of MMLE, common item equating can be accomplished via the computer program BILOG (Mislevy & Bock, 1990) as well as via the computer program MULTILOG (Thissen, 1991). In case of concurrent calibration of item parameter estimates using BILOG, the marginalization assumes there is a population distribution which may be either Gaussian or some arbitrary distribution jointly obtained with item parameter estimates (Bock & Aitkin, 1981; Mislevy & Bock, 1990). The appropriate specification of the population distribution has been shown to increase the accuracy of estimation (Seong, 1990). When there are two groups of examinees, MULTILOG default options calibrate items by constructing a unit normal metric for ability parameters of the base group. The mean ability of the target group is empirically obtained along with item parameters while fixing the standard deviation at unity. MULTILOG default options can be overridden so that the mean and the standard deviation of the target group also can be differently specified.

One unresolved issue in the context of concurrent calibration under MMLE is the form of the population ability distribution. (Note that the form of the population ability distribution is not an issue for JMLE.) In addition, there is a concern with appropriate specification of the target group population parameters. In a horizontal equating situation, these specifications do not normally cause serious problem as (1) the two distributions of abilities are

generally comparable and (2) the difficulty level of a well-designed test is typically matched to the ability of the examinee groups. In vertical equating, however, the specification becomes somewhat more complicated, particularly if the two ability distributions differ not only in location but also in variability.

Differences in simultaneous and concurrent calibration appear not to have been reported in the literature. In the present study, therefore, we compare linking and concurrent calibration. To illustrate the problems involved, in the following section, we present an illustrative example using three methods for developing a common metric: linking via a characteristic curve method (Stocking & Lord, 1983); concurrent calibration using marginal maximum a posteriori estimation as implemented in BILOG (Mislevy & Bock, 1990); and concurrent calibration using marginal maximum likelihood estimation as implemented in MULTILOG (Thissen, 1991). Following the example, we present results from a larger simulation study to study these issues in more detail.

Example

Data

We illustrate linking and concurrent calibration using data obtained from a standardized, multiple-choice university mathematics placement test. Examinees were entering freshmen at a large midwestern university who had not taken a college-level mathematics course. Originally there were three sections, A, B, and C, in the test. Examinees were told to take section B (intermediate and advanced algebra) and either section A (basic arithmetic, elementary algebra, and plane geometry) or section C (analytic

geometry and trigonometry) depending on their high school mathematics preparation. Examinees were advised to take section C if they had at least two-and-one-half years of high school mathematics (not including business mathematics). Otherwise, they were advised to take sections A and B. Consequently, students who took sections of B and C were better prepared in terms of their mathematics ability.

Two test data sets, set AB and set BC, were assembled using 10 items from each section. Form-AB consisted of 500 examinees' responses for the 20 items of sections A and B. Form-BC consisted of 500 examinees' responses for the 20 items of sections B and C. For purposes of this example, we refer to those who took Form-AB as the target group and those who took Form-BC as the base group. The task is to place the target groups examinees and item parameters onto the metric of the base group. The 10 items from section B were common to both tests.

Classical Item Statistics

Classical item difficulties, p_j, and biserial correlations between item score and item-excluded total score, r_j, are given in Table 1. Summary statistics for the separate Forms A, B, and C are presented in Table 2. For the common items, that is, Form B, note that the target group has a mean score of 3.182 whereas for the base group, the mean was 7.930. Clearly the common items were relatively easier for examinees in the base group.

Insert Tables 1 and 2 about here

Separate Calibrations Results

Both a two- and a three-parameter logistic model were fitted to the two data sets. The results, presented in Table 3, afford a comparison of the two models and indicate that addition of the asymptotic parameter did not significantly improve the fit of the model to the data. We therefore selected the simpler, two-parameter logistic model.

Insert Table 3 about here

Default options were used for the BILOG computer runs along with LOG, FRE, IDI=3, and RSC=3 options. The LOG option is used to place item and ability parameter estimates on the logistic metric (i.e., the scaling constant $D = 1.7$). The FRE option is used to estimate the population parameters of the underlying ability distribution jointly along with item parameters. The IDI=3 option uses the empirical prior from the item parameter estimation phase to estimate the examinee ability with the expected a posteriori (EAP) method. The RSC=3 option places the item parameter estimates from the target group onto the N(0,1) ability estimates metric of the target group. Item parameter estimates for both groups are reported in Table 4. The item parameter estimates of the base group were likewise expressed on the normalized estimated ability metric of the base group. It is important to note that the default estimation procedure of BILOG for the two-parameter model employs a lognormal prior for item discrimination. The estimation of item parameters, therefore, is marginal maximum a posteriori estimation (MMAPE).

Insert Table 4 about here

Item difficulty estimates of the target group were larger than those of the base group and suggest that the base group has higher ability. Unfortunately, such comparisons are premature as the items are not yet on a common metric.

Insert Tables 5 and 6 about here

In order to place item parameter estimates of the target group onto the metric of the base group, the test characteristic method of Stocking and Lord (1983) was used as implemented in the computer program EQUATE (Baker, 1993). The resulting linking coefficients, $A = .610$ and $B = -2.482$, were used to transform item parameter estimates of the target group to the metric of the base group (see Table 5). Note that parameter estimates of the common items, even after linking transformation, are not generally the same in the target and base groups. Hambleton and Swaminathan (1985) recommend averaging these estimates. One problem with averaging in this way is that, when the item parameter estimates are changed, the subsequent ability distribution of the base group may no longer be N(0,1). Summary statistics for linked item parameter estimates are given in Table 6. Using these same A and B coefficients, we can also express the ability estimates of the target group on the θ metric of the base group.

Concurrent Calibration Results Using BILOG

BILOG concurrent calibration results are reported in Table 7. For concurrent calibration, the 10 common items were treated as taken by both groups of

examinees. Items specific to the target or base group were treated as taken only by that group. Default options were again used for the BILOG runs with LOG, FRE, IDI=3, and RSC=3 options. Estimates of all 30 item parameters are placed by BILOG onto the N(0,1) estimated ability metric. BILOG estimates the underlying population ability distribution jointly along with item parameters.

Insert Tables 7 and 8 about here

As the concurrent standardization was based on the whole group of examinees, it is not yet appropriate to compare these item parameter estimates to those from the linked separate calibrations. We may use the characteristic curve method or the following procedure to make such comparison. First, we need to transform the concurrent calibration ability estimates of the base group to a N(0,1) metric. We next need to transform the item parameter estimates. The base group examinees had the mean of .758 of the ability estimates and the standard deviation of .783. We then make a transformation of the target group ability estimates to N(0,1). In this case we linearly reexpress the estimates onto the arbitrary metric of the standard normal of the ability estimates of the base group. The item parameter estimates of the rescaled group are also reported in Table 7. Table 8 contains summary statistics of the item parameter estimates from the BILOG concurrent calibration. It is clear that there are differences between the summary statistics from this concurrent calibration case and the linked separate calibrations case.

As noted earlier, one problem with concurrent calibration under MMLE is that of specifying the underlying ability distribution. In terms of

MMLE/MMAPE, the joint likelihood/posterior is marginalized under the assumption that a population distribution exists. In this example, even if we obtain this population distribution jointly along with the item parameter estimates, it is quite plausible to assume there might actually be two different forms of underlying ability distributions. If the target group population distribution of ability is truly different from that of the base group, then marginalization of the likelihood function under the assumption of a single ability distribution may not be the correct specification. There are two concerns in this regard. First, how and when can we be sure that there exists only one underlying distribution? Second, what is the effect on the subsequent metric of a misspecification of the ability distribution?

In part, concurrent calibration can potentially remove some equating errors which arise in the case of separate calibrations. It could possibly also remove some of the arbitrariness of the decisions made in linking. It should be noted, however, that the concurrent calibration may not always be either possible or economical. For example, item parameter estimates obtained on earlier forms of a test will generally differ to some extent from current estimates. Subsequent combination of existing data with new data just to achieve concurrent calibration results may also incur different equating errors.

Concurrent Calibration Results Using MULTILOG

MULTILOG employs a similar marginalization process to that used in the BILOG computer program. The main difference between BILOG and MULTILOG is that MULTILOG permits specification of different population ability parameters. This is particularly valuable if we have two or more groups of examinees. When two (or more) groups are calibrated, MULTILOG

BEST COPY AVAILABLE

assumes the base group's ability to be N(0,1) and the target group's ability to be N($\hat{\mu}$,1), where $\hat{\mu}$ is the estimated mean ability of the target group. The mean ability of the target group can be obtained jointly by MULTILOG along with item parameters.

Default options were used for estimation of item parameters. Under these conditions, MULTILOG provides MMLE estimates of item parameters. Concurrent calibration results using MULTILOG are reported in Table 9. For this estimation, the 10 common items were constrained to have the same estimates for both the target and base groups. Note that the item parameter estimates of the base group were placed on a N(0,1) metric. The target group's ability was estimated as N(−1.790,1). All item parameter were expressed on the base group's metric of N(0,1). Summary statistics are given in Table 10.

Insert Tables 9 and 10 about here

One problem with using concurrent calibration via MULTILOG concerns the specification of the distribution of ability used for marginalization. In this example, we used the program default options to obtain the mean ability of the target group jointly with item parameter estimates. If the target group population distribution of ability is different from that of the base group, then marginalization of the likelihood function under the assumption that there are two different ability distributions is more appropriate.

Comparison of Results

Correlations between the item difficulty parameters estimated separately and concurrently were relatively high (see Table 11): The correlation between

linking and concurrent calibration via BILOG was .966 and that between linking and concurrent calibration via MULTILOG was .985. Similarly, the correlation between the BILOG and MULTILOG calibration results was .988. High correlations indicate strong linear relationships between the various item difficulty estimates.

Insert Table 11 about here

The correlations between linking and concurrent item discrimination estimates were not as high: The correlation between linking and BILOG estimates was .763; that between linking and MULTILOG estimates was .632. The correlation between the two concurrent calibration results was somewhat higher ($r = .879$).

Test characteristic curves (TCCs) from the separate calibrations are plotted in Figure 1 along with the TCCs from concurrent calibration via BILOG and TCCs from concurrent calibration via MULTILOG. If the three different procedures of obtaining common metrics were without errors, all three TCC plots should be identical. Generally, the three patterns appear to be quite similar.

Insert Figures 1 and 2 about here

Figure 2 presents the line of relationship between observed scores under the three different methods. Again, if the three were without error, all curves should be identical. It is clear that there are differences due to the three methods. The discrepancies, in fact, arise from differences between the sets of item parameter estimates.

The example is informative and serves to illustrate common item linking and concurrent calibrations. It does not provide clear information, however, regarding the comparable quality of the linking and concurrent calibrations. This is because we do not know the true parameters and, consequently, the form of the true relationship between the different metrics and the tests. When dealing with real data, there is no satisfactory way to evaluate methods of constructing a common metric as no criterion yet exists against which to check the accuracy of the results obtained. We can only make such judgements when we know what the proper relationship is between the two sets of item parameters. Such a criterion is available, however, if we use generated data sets. In the next section, we present results from a simulation study.

Simulation Study

Data Generation

In this section we compare the three methods of obtaining a common metric under IRT in the context of a recovery study design in order to more closely examine the effects of each of the methods with respect to known item and ability parameters. Data for the simulation study were generated for 50 items and 500 examinees using the computer program GENIRV (Baker, 1988). The two-parameter model was employed to generate item response vectors. The 50 sets of item parameters (see Table 12) were originally reported by Lord (1968) for the three-parameter model. Subsequently, the estimates have been used by both McLaughlin and Drasgow (1987) and Cohen and Kim (1993) in the context of the two-parameter model.

Insert Table 12 about here

The simulation study consisted of a single base group of examinees with an ability distribution generated N(0,1). Two different target groups were also generated with ability distributions N(0,1) and N(1,1), respectively. A total of 150 data sets each with 50 items and 500 examinees were generated by changing the random number seed. This included 50 base group data sets, 50 target group data sets for N(0,1), and 50 target group data sets for N(1,1).

Number of Common Items and Item Parameter Estimation

For each combination of a target group and the base group, four different lengths of common items sets were used: 5, 10, 25, and 50 items. For the 5-common item condition, items 1–5 in of Table 12 were used. For the 10-common item condition, items 1–10 of Table 12 were used. For the 25-common item condition, items 1–25 were assumed to be common items. The 50 common item condition simulated a typical differential item functioning detection situation in which all of the items need to be placed onto the same metric before comparisons could be made. The summary statistics of the item parameters of the four sets of the common items are reported in Table 13.

Insert Table 13 about here

For the separate calibrations, the computer program BILOG was used to estimate item parameters using default options along with FRE, IDI=3, and

RSC=3 options. First, base group item parameters were estimated and then target group item parameters were estimated. A total of 150 sets of item parameter estimates were obtained in this way from the total of 150 BILOG calibrations. Since we had four different linking situations corresponding to the four lengths of common item sets, for each combination of the base group and the target group, four EQUATE runs were performed. In case of the 5-common item condition, the EQUATE run produced linking coefficients A and B based on these 5 items. Then using A and B, item parameter estimates from the target group were placed onto the metric of the base group. Finally, the item parameter estimates from the common items were averaged to obtain the linked item parameter estimates. For the 5-common item condition, this resulted in estimates of item parameters for 95 items after the linking. A total of 400 EQUATE runs were performed, that is, 50 replications for the four EQUATE runs of the base group and the N(0,1) target group as well as for the four EQUATE runs of the base group and the target group of the N(1,1) target group.

For the concurrent calibrations, both BILOG and MULTILOG were used. First, 100 combined data sets were formed of the base group and each of the two target groups. Note that only two groups were analyzed on each concurrent run: the base group and the N(0,1) target group or the base group and the N(1,1) target group. A single combined data set was analyzed four times using BILOG and another four times using MULTILOG. The four computer runs were performed for each of four common item conditions. Altogether, 400 BILOG runs were performed and 400 MULTILOG runs were performed.

For the BILOG runs, default options were employed along with FRE, IDI=3, and RSC=3 options resulting in MMAPE estimation. The final item

and ability estimates were expressed on the N(0,1) metric of the estimated ability parameters. For the MULTILOG runs, all default options were used and the item parameter estimation was MMLE.

Equating and Evaluation Criteria

The final estimates for item and ability parameters from the separate calibrations used for the linking simulations were all expressed on the N(0,1) metric of the base group ability estimates.

Estimates of item and ability parameters used for the concurrent calibrations were placed by BILOG onto the metric of combined ability estimates. For the base and target group combination with the same N(0,1) ability distributions, the final estimates were placed on the same metric of the generating parameters. In case of the combination of the base group of N(0,1) and the target group of N(1,1), the resulting metric was based on the standardized metric of the combined ability parameter estimates. The metrics from the concurrent calibrations using MULTILOG was based only on the base group metric of N(0,1).

These final estimates from separate linking calibrations and concurrent calibrations are not directly comparable. In order to make comparisons of the estimates, additional EQUATE runs were performed to place all item parameter estimates onto the metric of generating item parameters. In case of the 5-common item condition, 95 sets of common items were equated to the metric of generated item parameters. In case of the 10 common item condition, 90 sets of common items were equated back to the metric of the generated item parameters. All together, 1,200 EQUATE runs were required to place final estimates onto a common metric.

One means of evaluating results from the different methods of obtaining

a common metric is to compare equating coefficients to expected values. A more definitive description is possible, however, in a recovery study. Root mean square differences (RMSDs) between the estimates and the generating parameters provide a good indication of the quality of the recovery and, thereby, an indication of the quality of linking and concurrent calibrations. The smaller the RMSDs, the better the methods of obtaining a common metric. RMSDs were calculated separately for each parameter, once for item discrimination and once for item difficulty. The RMSD for item discrimination is defined as

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(a_j - \alpha_j)^2}, \tag{1}$$

where n is the total number of items. Recall that the total number of items were 95, 90, 75, and 50 for each common item condition of 5, 10, 75, and 50 items, respectively. Note that the item parameter estimates for both separate and concurrent calibrations were linked back to the metric of the generating item parameters before calculating the RMSDs. For item difficulty, the RMSD is defined as

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(b_j - \beta_j)^2}. \tag{2}$$

Since it is possible that a method of obtaining a common metric may function better at recovery of one type of item parameter than at recovery of the other, it is also useful to consider a single index which could describe simultaneously the quality of the recovery for both parameters. The mean Euclidean distance (MED) provides such an index. The MED is the average of the square roots of the sum of the squared differences between the discrimination and difficulty parameter estimates and their generating values.

The MED is defined as

$$\frac{1}{n}\sum_{j=1}^{n}\sqrt{(\hat{\xi}_j-\xi_j)'(\hat{\xi}_j-\xi_j)}, \tag{3}$$

where $\hat{\xi}_j = (a_j, b_j)'$ and $\xi_j = (\alpha_j, \beta_j)'$. MEDs were calculated between the underlying parameters and their estimates. One caveat in using the MED, of course, is that item discrimination and difficulty parameters are not expressed in comparable and interchangeable metrics. Even so, the MED does provide a potentially useful descriptive index.

Results

Root Mean Square Differences

Recovery of the underlying parameters was first evaluated with RMSDs between the transformed estimates and the generating parameters for each method for obtaining a common metric. The results for item discrimination, summarized in Table 14, indicate that the separate calibrations condition yielded generally smaller RMSDs for item discrimination (see also Figure 3). For the N(1,1) target group case, both separate calibrations and the concurrent calibration via BILOG yielded similar results except for the 5-common items condition. The concurrent calibration via MULTILOG yielded larger RMSDs of item discrimination across all conditions.

Insert Table 14 and Figure 3 about here

RMSDs for item discrimination for the cases in which the N(1,1) target group was used were smaller than cases in which the N(0,1) target group was used. As can be seen in Table 14, there is a clear tendency for the sizes of

the RMSDs for item discrimination to decrease as the number of common items increased.

RMSDs for item difficulty are also reported in Table 14. Concurrent calibration via BILOG yielded smaller RMSDs for the N(1,0) target group condition. As the number of common items increased under the N(1,0) target group condition, all three methods yielded essentially the same results. For the N(1,1) target group condition, the separate calibrations yielded slightly smaller RMSDs except for the 50-common item case. Concurrent calibration via BILOG yielded somewhat larger RMSDs for both the 5- and 10-common item conditions. There did not appear to be any systematic relationship between the distribution of the target group's ability and the size of the item difficulty RMSDs. For both groups, as the number of common items increased, the size of the RMSDs of item difficulty decreased.

Mean Euclidean Distances

Trends for MEDs between item parameter estimates and underlying parameters were similar to those reported for RMSDs. Table 15 and Figure 4 present the MED results. Linking yielded smaller MEDs for all conditions except the N(1,1) target group with 50 common items. Concurrent calibration via MULTILOG yielded larger MEDs under the N(0,1) target ability condition. BILOG concurrent calibration produced larger MEDs for the N(1,1) target ability condition with 5 common items. The size of the average MEDs decreased as the number of common items increased. Also differences among methods of obtaining a common metric decreased as the sizes of common items increased.

Insert Table 15 and Figure 4 about here

Linking Coefficients and Population Parameter Estimates

As noted above, transformation of item parameter estimates from the target group onto the base group metric was accomplished using A and B equating coefficients. In terms of separate calibration, therefore, it was of interest to look at the values of these coefficients as we know apriori the theoretically expected values. The theoretically expected values of A and B for placing the target group ability of N(0,1) onto the base group ability of N(0,1) are 1 and 0, respectively. For placing the N(1,1) target group onto the N(1,0) base group metric, the values of A and B are 1 and 1, respectively. Summary statistics of the equating coefficients from the separate calibrations for two different target group ability distributions and four numbers of common items are reported in Table 16.

Insert Table 16 about here

Differences in equating coefficients from expected values were generally small for all simulated conditions. For the N(0,1) target group, the A and B were essentially 1 and 0 for all common item conditions. Likewise, for the N(1,1) target group, the A and B were essentially 1 and 1 for all common item conditions.

In case of the concurrent calibration via MULTILOG, the base group ability metric was set to N(0,1). The mean ability of the target group (i.e., population parameter or hyperparameter) was jointly estimated along with the item parameters. Standard deviations of ability for the base group and

the target group were both fixed at 1. For the N(0,1) target group, the expected population mean was 0 and for the N(1,1) target group it was 1.

Insert Table 17 about here

Table 17 contains means and standard deviations of the population parameter estimates over fifty replications for both target group ability conditions and for the 4 common item conditions. The hyperparameter of the target group N(0,1) (i.e., the posterior population mean) was not close to the expected value. All values were smaller than the expected value of 0. As the number of common items increased, the mean of the hyperparameters approached but did not reach the theoretically expected value. The mean hyperparameter for the N(1,1) target group was also less than the expected value of 1. As was seen for the N(0,1) target group case, as the number of common items increased, the mean hyperparameters tended to approach a value of 1.

Summary and Discussion

The comparability of IRT item parameter estimates across different tests measuring the same underlying ability is an important matter for test developers and researchers since all decisions about examinees are derived from these estimates. A number of different methods are available for developing common metrics, not all of which yield the same ability estimates. Which method to choose to develop a common metric is often a matter of uncertainty and concern. In this paper, we have presented examples and simulation study results using three different, commonly used methods for obtaining a common metric in IRT. The three methods were linking

of separately calibrated metrics using linear equating coefficients A and B obtained from the test characteristic curve method, concurrent calibration via MMAPE, and concurrent calibration via MMLE.

In the recovery study section of this paper, comparisons were made of the similarities between generating parameters and item parameter estimates obtained after transformation of the results from each of the methods to the underlying metric. The simulation results indicated that recovery for linking of separate calibrations was generally better than recovery from either of the concurrent calibrations. The finding of greatest interest was that, when the ability of the base group was not well-matched to the ability of the target group and when small numbers of common items were used, concurrent calibration via BILOG resulted in somewhat larger RMSDs and MEDs. In addition, for concurrent calibration via MULTILOG with the N(0,1) target group, both RMSDs and MEDs were somewhat larger. As the number of common items increased, however, all three methods tended to yield similar results. Comparisons of the three methods in terms of item parameter estimation and ability estimation are presented in Table 18.

Insert Table 18 about here

Differences among the methods compared in this study were primarily ones inherent to the indeterminacy of the IRT ability metric. It is well-known that the ability metric in IRT is unique up to a linear transformation. Both linking and concurrent calibration are closely related to the problem of the metric indeterminacy.

Computer programs for estimating item and ability parameters under IRT resolve the problem of the linear indeterminacy of the metric problem in

different ways. LOGIST, for example, resolves this problem by standardizing the estimates of ability so that the estimates have a mean of 0 and a standard deviation of 1. In BILOG and MULTILOG, the ability parameters are not estimated with item parameters. Consequently, when comparing the results from the BILOG and MULTILOG programs, care needs to be exercised to deal within the frame of reference of the metric of the item and ability estimates. With respect to BILOG, at the end of the item parameter estimation phase, the estimated posterior ability distribution is used to establish the metric of item parameter estimates. Mislevy and Bock (1990, p. 1-19) also recommend use of the FRE option for common item equating. The resulting estimated posterior ability distribution is based on the discrete distribution over a finite number of points (Mislevy, 1984). BILOG ability parameters in this study were estimated and normalized.

Mislevy and Stocking (1989) recommend use of the expected a posteriori (EAP) for estimation with the empirical examinee ability distribution during the item parameter estimation phase. Therefore, for BILOG runs in the present study, the EAP method was used to estimate ability parameters. Again, the empirical prior distribution estimated during the item parameter estimation phase (i.e., phase 2 of BILOG) was used as the designated type of prior distribution for scale scores (IDI=3). With the RSC=3 option, the ability estimates were transformed onto a standard normal metric, that is, $\hat{\theta} \sim N(0,1)$. Item parameter estimates were then expressed on this metric.

For MULTILOG runs in this study, ability parameters were not estimated. The normalized posterior of the base group latent ability distribution provides the underlying metric in MULTILOG. When comparing linking results (i.e., for separate calibrations), the MULTILOG item parameter estimates, therefore, cannot be viewed in the same manner as BILOG

BEST COPY AVAILABLE

results. The differences in the metrics of BILOG and MULTILOG results rest fundamentally on the fact that the posterior ability distribution from the item parameter estimation phase under MMLE is not the same as the empirical distribution of the estimates of ability. Mislevy (1984) has shown that the estimated distribution of ability is not the same as the empirical distribution of ability estimates.

It is possible to use MULTILOG to compute maximum likelihood or maximum a posteriori estimates of ability. If this had been done, then it would also have been possible to calculate the mean of ability estimates for the target group and the base group separately. The ability estimates of the base group could then be normalized followed by an additional transformation of the ability estimates of the target group to the metric of the base group in order to permit comparisons between the underlying population mean and the mean of the transformed estimates of the target group.

One of the factors playing a role in determining the metric for both BILOG and MULTILOG is the form of the prior distribution imposed on the item discrimination parameters. Under the two-parameter model, BILOG default options place a lognormal prior distribution on the item discrimination parameter. The estimation in this study provided by BILOG was MMAPE whereas that by MULTILOG was MMLE.

As the results of this study indicate, the scales resulting from the three different methods were not the same. Therefore, it was necessary to perform an additional linking of the linked item parameter estimates from the separate calibrations and of the concurrent calibration results from both BILOG and MULTILOG to the underlying metric before RMSDs and MEDs could be obtained. A linear transformation, such as that due to Stocking and Lord

(1983) and the one used in this study, puts item parameter estimates onto the metric of underlying parameters. Remaining differences between estimates and parameters can be attributed to estimation errors.

It is well-known that estimation of the pseudo-guessing parameter in the three-parameter model can be problematic under some circumstances (e.g., small samples). One solution in this regard is to fix these parameters to some value before estimating item and ability parameters. In terms of the recovery study, the pseudo-guessing can always be fixed to the generating values. We have no such luxury, however, for real data. For purposes of either linking or equating, the final estimate of the pseudo-guessing parameter used for an item should be equal for both groups of examinees. A number of ways exist to obtain these values. One way is to use the average values from the initial calibrations. This may be less desirable if there are marked differences in ability distributions between groups. In such cases, it may be more appropriate to use estimates from the group with the lower ability, particularly as this might afford greater information for estimating the pseudo-guessing parameters. It is also possible to use external information to fix all the pseudo-guessing parameters. One approach is to set the value based on the chance probability of a correct response given the number of alternatives. Empirical Bayes estimation may also be useful, particularly in those cases for which sufficient information is not available at the lower end of the ability distribution. When using BILOG, once the pseudo-guessing estimates have been obtained from the respective target and base groups, it is then possible to perform additional calibrations to obtain the item parameter estimates while fixing the pseudo-guessing parameters. For such cases, however, it is often necessary to use relatively strong priors.

Results from the present study suggest that, in general, when the number of common items is small, linking of separate calibrations may be preferable to concurrent calibration. Further, when the number of common items is large, both types of procedures appear to function similarly. When estimation algorithms such as MMAPE or MMLE are used, care needs to be taken in the proper specification of the population ability distributions involved. Further studies are needed of methods for obtaining a common metric under IRT and of the impact of prior assumptions on the resulting θ metric.

References

Baker, F. B. (1988). *GENIRV: A program to generate item response vectors* [Computer program]. Madison, University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement*, *17*, 20.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Cohen, A. S., & Kim, S.-H. (1993). A comparison of Lord's χ^2 and Raju's area measures in detection of DIF. *Applied Psychological Measurement*, *17*, 39–52.

Divgi, D. R. (1980, April). *Evaluation of scales for multilevel test batteries*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement*, *9*, 413–415.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Linn, R. L., Levine, M. V., Hasting, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comparison. *Applied Psychological Measurement*, *5*, 159–173.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*, 989–1020.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 169–194.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139–160.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement*, *11*, 161–173.

Mediax Interactive Technologies (1985). *MICROSCALE* [Computer program]. West Port, CT: Author.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57–75.

Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*, 299–311.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*, 333–344.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.

Table 1

Classical Item Statistics from Target Group and Base Group

	Target Group		Base Group		
Item	p_j	r_j	p_j	r_j	Form
1	.784	.254			A
2	.828	.221			A
3	.670	.346			A
4	.664	.454			A
5	.624	.353			A
6	.658	.299			A
7	.728	.518			A
8	.652	.306			A
9	.382	.302			A
10	.830	.428			A
11	.254	.288	.870	.473	B
12	.402	.200	.888	.513	B
13	.318	.170	.852	.356	B
14	.166	.222	.806	.672	B
15	.496	.352	.870	.353	B
16	.500	.282	.838	.155	B
17	.374	.173	.748	.476	B
18	.192	.428	.610	.439	B
19	.338	.291	.790	.514	B
20	.142	.150	.658	.426	B
21			.718	.426	C
22			.850	.475	C
23			.592	.504	C
24			.664	.481	C
25			.358	.453	C
26			.666	.459	C
27			.638	.428	C
28			.688	.441	C
29			.890	.573	C
30			.440	.521	C
	$N = 500$		$N = 500$		

Table 2

Summary Statistics for Forms A, B, and C

	Target Group			Base Group			
Form	Mean	SD	Alpha[a]	Mean	SD	Alpha	Item
A	6.820	2.011	.556				1–10
B	3.182	1.765	.389	7.930	1.811	.581	11–20
C				6.504	2.250	.663	21–30

[a]Cronbach's Alpha.

Table 3

Difference of the Marginal Log Likelihoods

	Target Group	Base Group	Number of Item
Model	2 Log Likelihood	2 Log Likelihood	Parameters
3PM	−11394.0079	−9929.8318	60
2PM	−11409.6836	−9946.8785	40
Difference	15.6757	17.0467	20

Table 4

Item Parameter Estimates of the Two-Parameter Logistic Model

	Target Group		Base Group	
Item	a_j	b_j	a_j	b_j
1	.675	−2.087		
2	.607	−2.776		
3	.792	−1.018		
4	1.115	−.768		
5	.850	−.694		
6	.704	−1.034		
7	1.518	−.916		
8	.705	−.991		
9	.717	.739		
10	1.252	−1.607		
11	.762	1.574	1.176	−1.978
12	.504	.828	1.367	−1.939
13	.501	1.602	.855	−2.314
14	.614	2.814	2.074	−1.127
15	.852	.013	.786	−2.678
16	.690	−.007	.498	−3.459
17	.459	1.170	1.162	−1.185
18	1.067	1.621	1.014	−.553
19	.695	1.060	1.337	−1.308
20	.498	3.776	.970	−.820
21			1.007	−1.126
22			1.138	−1.859
23			1.157	−.432
24			1.184	−.753
25			1.042	.670
26			1.093	−.799
27			.974	−.712
28			1.011	−.957
29			1.636	−1.761
30			1.316	.215
	$\theta_{\text{Target}} \sim \text{N}(0,1)$		$\theta_{\text{Base}} \sim \text{N}(0,1)$	

Table 5

Linked Item Parameter Estimates of the Two-Parameter Logistic Model

	Target Group[a]		Base Group		Linked Group	
Item	a_j	b_j	a_j	b_j	a_j	b_j
1	1.105	−3.756			1.105	−3.756
2	.994	−4.177			.994	−4.177
3	1.297	−3.103			1.297	−3.103
4	1.827	−2.951			1.827	−2.951
5	1.392	−2.906			1.392	−2.906
6	1.153	−3.113			1.153	−3.113
7	2.487	−3.041			2.487	−3.041
8	1.156	−3.087			1.156	−3.087
9	1.174	−2.031			1.174	−2.031
10	2.051	−3.463			2.051	−3.463
11	1.249	−1.521	1.176	−1.978	1.213	−1.749
12	.826	−1.977	1.367	−1.939	1.096	−1.958
13	.821	−1.504	.855	−2.314	.838	−1.909
14	1.006	−.764	2.074	−1.127	1.540	−.946
15	1.396	−2.474	.786	−2.678	1.091	−2.576
16	1.130	−2.486	.498	−3.459	.814	−2.972
17	.752	−1.767	1.162	−1.185	.957	−1.476
18	1.747	−1.492	1.014	−.553	1.381	−1.023
19	1.139	−1.835	1.337	−1.308	1.238	−1.571
20	.816	−.177	.970	−.820	.893	−.498
21			1.007	−1.126	1.007	−1.126
22			1.138	−1.859	1.138	−1.859
23			1.157	−.432	1.157	−.432
24			1.184	−.753	1.184	−.753
25			1.042	.670	1.042	.670
26			1.093	−.799	1.093	−.799
27			.974	−.712	.974	−.712
28			1.011	−.957	1.011	−.957
29			1.636	−1.761	1.636	−1.761
30			1.316	.215	1.316	.215

[a]Linking coefficients are $A = .610$ and $B = -2.482$.

Table 6

Summary Statistics of Linked Item Parameter Estimates

	Linked Group			
	a_j		b_j	
Item	Mean	SD	Mean	SD
1–10	1.464	.492	−3.163	.567
11–20	1.106	.240	−1.668	.747
21–30	1.156	.197	−.751	.781
Total	1.242	.362	−1.861	1.218

Table 7

Item Parameter Estimates from Concurrent Calibration via BILOG

	Combined Group[a]		Rescaled Group	
Item	a_j	b_j	a_j	b_j
1	.994	−2.090	.778	−3.635
2	.918	−2.510	.719	−4.172
3	1.328	−1.310	1.041	−2.640
4	1.924	−1.153	1.507	−2.440
5	1.474	−1.110	1.155	−2.384
6	1.126	−1.345	.882	−2.685
7	2.501	−1.249	1.959	−2.562
8	1.086	−1.339	.851	−2.677
9	1.096	−.239	.859	−1.273
10	2.006	−1.672	1.571	−3.102
11	2.068	−.273	1.620	−1.316
12	1.643	−.585	1.287	−1.715
13	1.505	−.370	1.179	−1.439
14	2.418	−.029	1.894	−1.004
15	1.422	−.776	1.114	−1.959
16	1.042	−.845	.817	−2.046
17	1.175	−.297	.920	−1.347
18	1.456	.356	1.141	−.513
19	1.571	−.286	1.231	−1.333
20	1.537	.347	1.204	−.525
21	1.126	−.256	.882	−1.295
22	1.236	−.942	.969	−2.171
23	1.341	.371	1.050	−.495
24	1.278	.065	1.001	−.885
25	1.222	1.304	.957	.697
26	1.228	.038	.962	−.919
27	1.074	.104	.841	−.835
28	1.115	−.114	.873	−1.113
29	1.701	−.886	1.332	−2.099
30	1.547	.925	1.212	.213
	$\theta_{\text{Combined}} \sim \text{N}(0,1)$			

[a]The base group is $\text{N}(.758, .783^2)$.

Table 8

Summary Statistics of Item Parameter Estimates from BILOG

	Combined Group				Rescaled Group			
	a_j		b_j		a_j		b_j	
Item	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1–10	1.445	.528	−1.402	.607	1.132	.413	−2.757	.774
11–20	1.584	.401	−.276	.412	1.241	.314	−1.320	.526
21–30	1.287	.198	.061	.701	1.008	.155	−.890	.895
Total	1.439	.405	−.539	.851	1.127	.317	−1.656	1.086

Table 9

Item Parameter Estimates from Concurrent Calibration via MULTILOG

	Target Group		Base Group	
Item	a_j	b_j	a_j	b_j
1	.824	−3.470		
2	.770	−3.950		
3	1.010	−2.570		
4	1.260	−2.430		
5	1.050	−2.330		
6	.876	−2.600		
7	1.640	−2.570		
8	.853	−2.590		
9	.826	−1.150		
10	1.390	−3.170		
11	1.530	−1.060	1.530	−1.060
12	1.190	−1.510	1.190	−1.510
13	1.140	−1.200	1.140	−1.200
14	1.720	−.706	1.720	−.706
15	1.010	−1.790	1.010	−1.790
16	.782	−1.850	.782	−1.850
17	.845	−1.120	.845	−1.120
18	1.080	−.217	1.080	−.217
19	1.120	−1.090	1.120	−1.090
20	1.190	−.236	1.190	−.236
21			1.070	−.715
22			1.200	−1.420
23			1.250	−.058
24			1.210	−.381
25			1.170	.935
26			1.160	−.409
27			1.040	−.328
28			1.060	−.565
29			1.490	−1.480
30			1.410	.543
$\theta_{\text{Target}} \sim \text{N}(-1.79, 1)$			$\theta_{\text{Base}} \sim \text{N}(0,1)$	

Table 10

Summary Statistics of Item Parameter Estimates from MULTILOG

	MULTILOG			
	a_j		b_j	
Item	Mean	SD	Mean	SD
1–10	1.050	.290	−2.683	.748
11–20	1.161	.284	−1.078	.567
21–30	1.206	.147	−.388	.755
Total	1.139	.250	−1.383	1.186

Table 11
Correlations of Item Parameter Estimates

				Concurrent Calibration			
		Linking		BILOG		MULTILOG	
Method	Estimate	a_j	b_j	a_j	b_j	a_j	b_j
Linking	a_j	1.000					
	b_j	−.273	1.000				
BILOG	a_j	.763	.003	1.000			
	b_j	−.264	.966	.037	1.000		
MULTILOG	a_j	.632	.337	.879	.292	1.000	
	b_j	−.308	.985	−.015	.988	.289	1.000

Table 12
Item Parameters

Item	α_j	β_j
1	1.1	−.7
2	.7	−.6
3	1.4	.1
4	.9	.9
5	1.2	.7
6	1.6	1.1
7	1.6	1.1
8	1.6	−.1
9	1.2	.5
10	2.0	1.6
11	1.0	1.6
12	1.5	1.7
13	1.0	.7
14	1.1	2.0
15	1.1	2.4
16	2.0	1.4
17	1.7	1.3
18	.5	−.6
19	.9	1.6
20	1.3	.4
21	1.1	1.2
22	1.2	1.1
23	1.3	.2
24	1.3	.2
25	.5	−.8
26	.7	.5
27	.7	.5
28	.4	−.4
29	.4	−.4
30	1.2	−.5
31	.7	−1.0
32	.7	−.2
33	.7	−.2
34	.5	.0
35	.9	.5
36	1.1	1.4
37	1.2	−.6
38	1.2	−.6
39	.6	−.5
40	1.6	.3
41	1.1	.0
42	1.5	2.0
43	1.9	1.9
44	.9	−.5
45	.7	−.5
46	1.4	1.6
47	1.4	1.6
48	1.0	1.7
49	1.2	1.1
50	1.2	1.1

Table 13

Summary Statistics of Item Parameters of Common Items

	α_j			β_j		
Item	Mean	SD	Range	Mean	SD	Range
1–5	1.060	.270	[.7, 1.4]	.080	.729	[−.7, .9]
1–10	1.330	.386	[.7, 2.0]	.460	.766	[−.7, 1.6]
1–25	1.232	.391	[.5, 2.0]	.760	.884	[−.8, 2.4]
1–50	1.114	.405	[.4, 2.0]	.556	.923	[−1.0, 2.4]

Table 14

Mean and Standard Deviation of Root Mean Square Differences over Fifty Replications

Target Ability[a]	n_c^b	Separate Calibrations				Concurrent BILOG				Concurrent MULTILOG			
		Discrimination		Difficulty		Discrimination		Difficulty		Discrimination		Difficulty	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
N(0,1)	5	.154	.022	.110	.012	.164	.021	.102	.008	.211	.053	.127	.015
	10	.145	.023	.104	.009	.160	.023	.102	.009	.197	.038	.112	.010
	25	.130	.023	.094	.009	.151	.024	.094	.009	.180	.035	.098	.010
	50	.108	.018	.073	.008	.115	.019	.073	.009	.144	.025	.075	.009
N(1,1)	5	.145	.019	.111	.011	.153	.015	.251	.015	.173	.033	.123	.017
	10	.135	.019	.108	.011	.132	.015	.128	.014	.162	.025	.113	.015
	25	.121	.018	.101	.012	.122	.017	.104	.013	.145	.025	.103	.016
	50	.096	.012	.074	.010	.092	.013	.066	.008	.102	.015	.066	.009

[a] Base ability is N(0,1).

[b] Number of Common Items

Table 15

Mean and Standard Deviation of Mean Euclidean Distances over Fifty Replications

Target Ability[a]	n_c^b	Separate Calibrations		Concurrent BILOG		Concurrent MULTILOG	
		Mean	SD	Mean	SD	Mean	SD
N(0,1)	5	.152	.013	.151	.011	.186	.017
	10	.143	.010	.149	.011	.169	.014
	25	.129	.010	.139	.012	.152	.013
	50	.105	.010	.107	.011	.123	.013
N(1,1)	5	.149	.012	.269	.015	.168	.012
	10	.141	.011	.154	.011	.153	.012
	25	.129	.010	.130	.009	.137	.011
	50	.102	.009	.095	.008	.099	.009

[a]Base ability is N(0,1).

[b]Number of Common Items

Table 16

Mean and Standard Deviation of Equating Coefficients over Fifty Replications

Target Ability[a]	n_c^b	Coefficient A		Coefficient B	
		Mean	SD	Mean	SD
N(0,1)	5	.987	.064	.000	.050
	10	.994	.042	.007	.041
	25	1.000	.035	.005	.032
	50	.997	.025	.007	.022
N(1,1)	5	1.038	.065	1.048	.062
	10	1.026	.042	1.052	.036
	25	1.010	.040	1.041	.028
	50	1.014	.029	1.037	.026

[a]Base ability is N(0,1).

[b]Number of Common Items

Table 17

Mean and Standard Deviation of $\hat{\mu}$ from MULTILOG over Fifty Replications

Target Ability[a]	n_c^b	Population Parameter $\hat{\mu}$ Mean	SD
N(0,1)	5	−.452	.030
	10	−.417	.023
	25	−.400	.020
	50	−.385	.016
N(1,1)	5	.524	.031
	10	.554	.027
	25	.566	.025
	50	.573	.025

[a]Base ability is N(0,1).

[b]Number of Common Items

Table 18
Comparison of the Methods of Developing a Common Metric

	Method			
	Separate Calibrations and Linking		Concurrent Calibration	
			via BILOG	via MULTILOG
Group Analyzed	Target	Base	Combined	Combined
Computer Program	BILOG	BILOG	BILOG	MULTILOG
Item Parameter Estimation	MMAPE	MMAPE	MMAPE	MMLE
Latent Ability Distribution	Empirical	Empirical	Empirical	$\theta_{\text{Target}} \sim N(\mu,1)$, $\theta_{\text{Base}} \sim N(0,1)$
Prior on Item Parameter	$\log a_j \sim N(0,.5^2)$	$\log a_j \sim N(0,.5^2)$	$\log a_j \sim N(0,.5^2)$	None
Ability Estimation	EAP	EAP	EAP	NA
Prior on Ability	Empirical	Empirical	Empirical	NA
Metric	$\hat{\theta}_{\text{Target}} \sim N(0,1)$	$\hat{\theta}_{\text{Base}} \sim N(0,1)$	$\hat{\theta}_{\text{Combined}} \sim N(0,1)$	$\theta_{\text{Base}} \sim N(0,1)$
Linking Method	TCC via EQUATE		None	None
Final Metric	$\hat{\theta}_{\text{Base}} \sim N(0,1)$		$\hat{\theta}_{\text{Combined}} \sim N(0,1)$	$\theta_{\text{Base}} \sim N(0,1)$
Linking to Underlying Parameter	TCC		TCC	TCC

Figure Captions

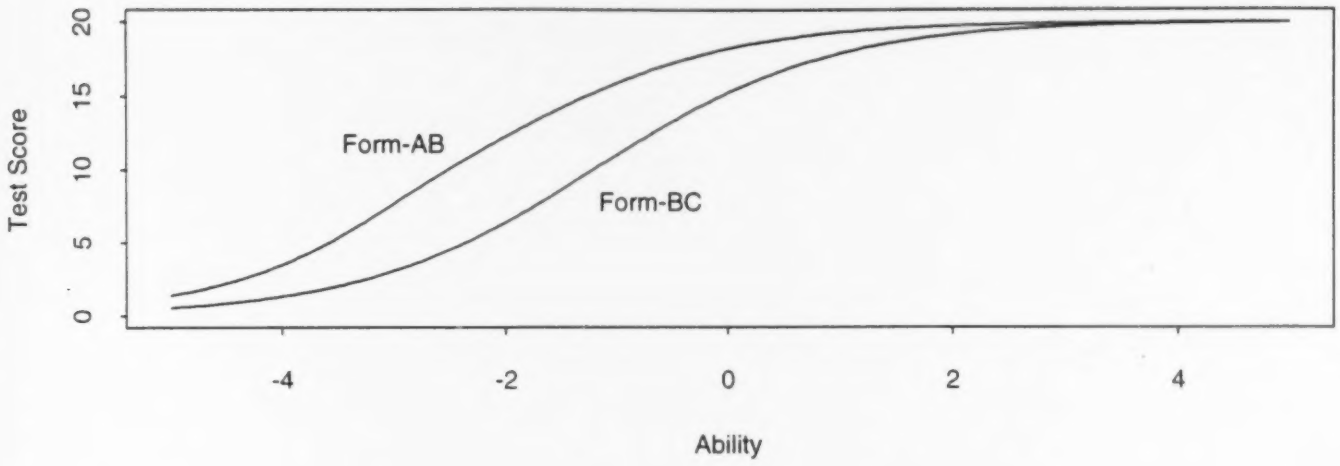
Figure 1. Test characteristic curves.

Figure 2. Line of relationship between two forms.

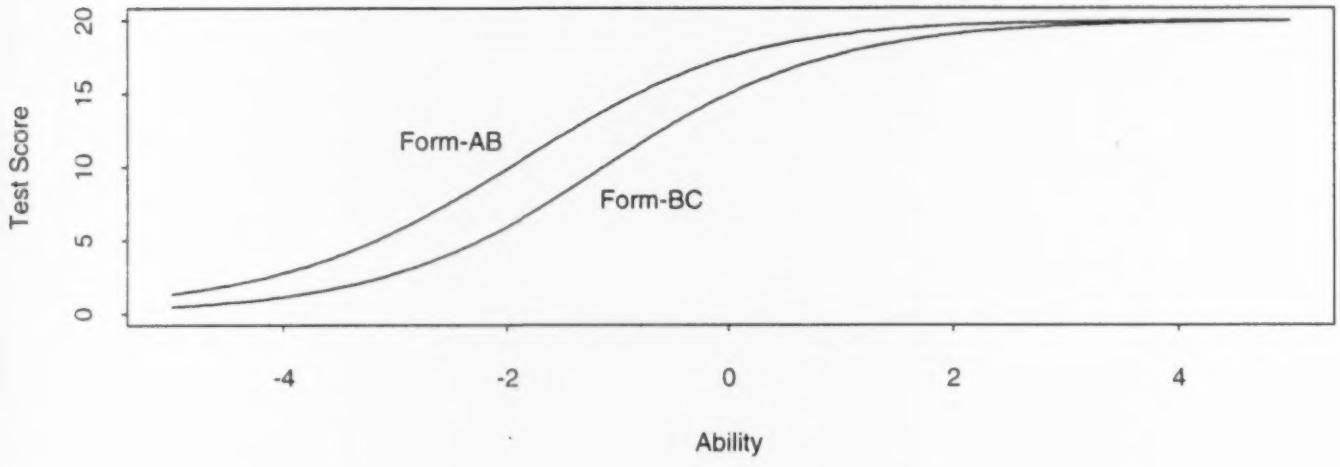
Figure 3. Root mean square differences results.

Figure 4. Mean Euclidean distance results.

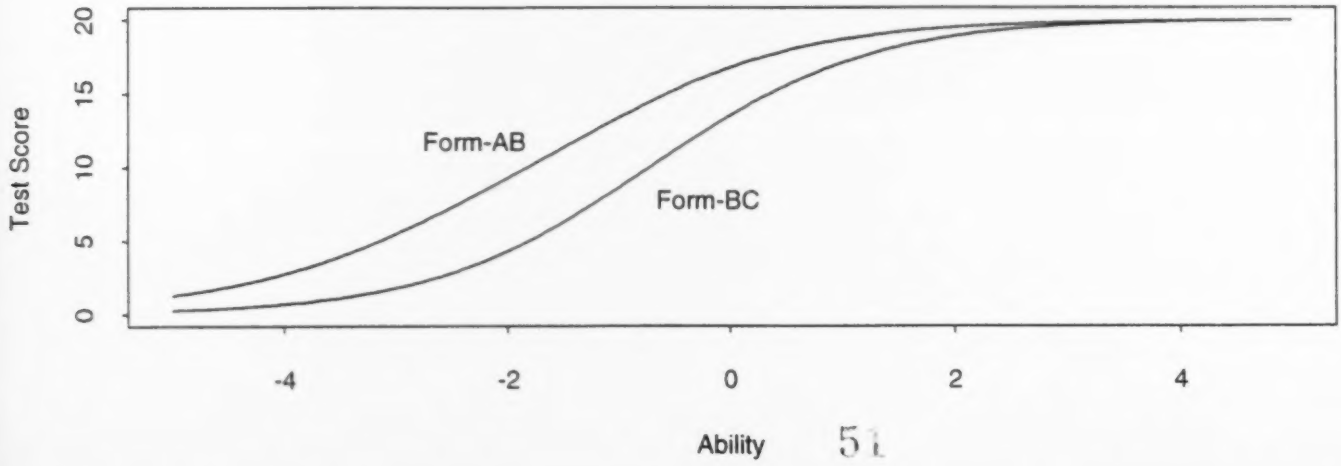
TCCs from Separate Calibrations



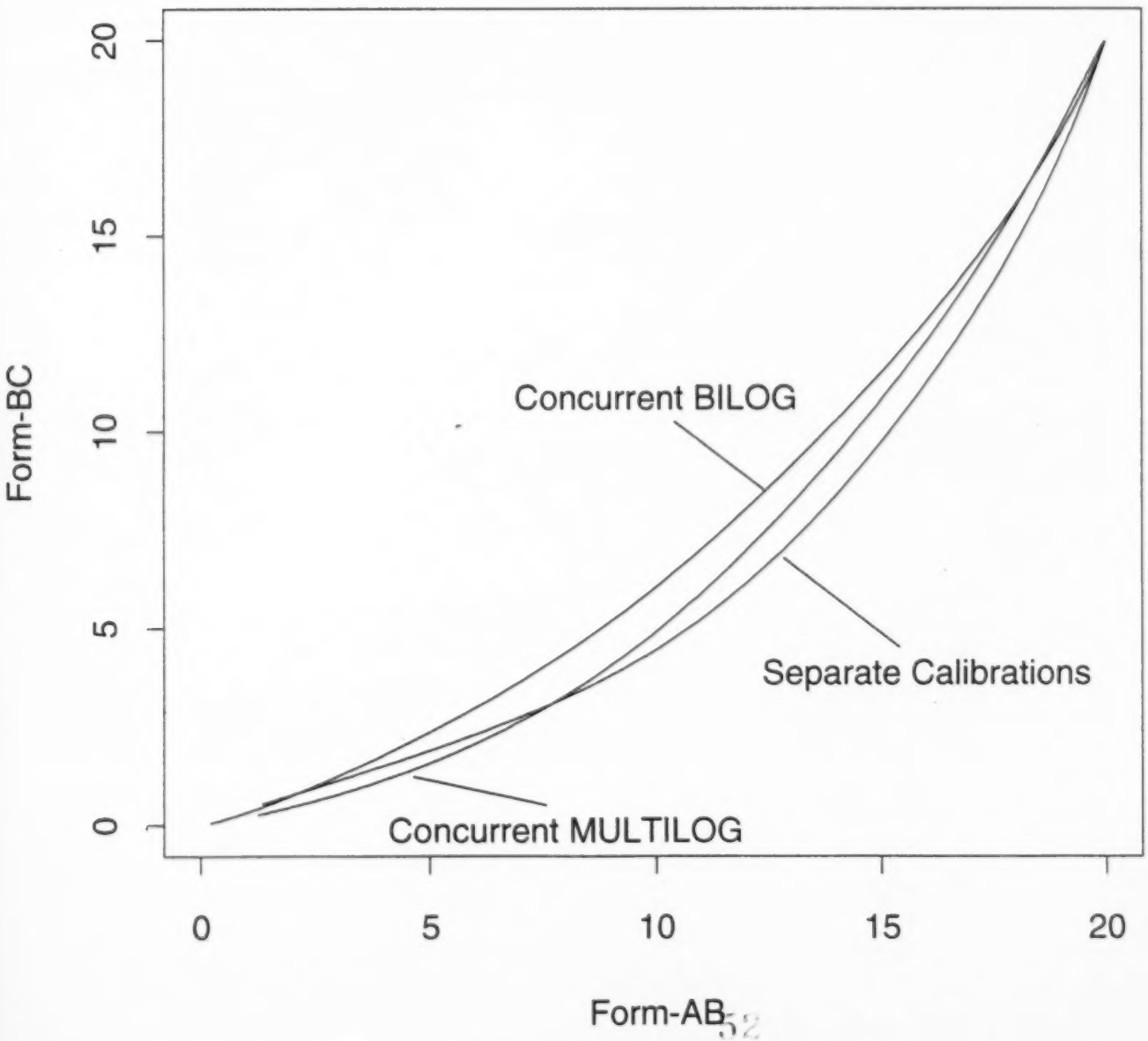
TCCs from BILOG Concurrent Calibrations



TCCs from MULTILOG Concurrent Calibrations

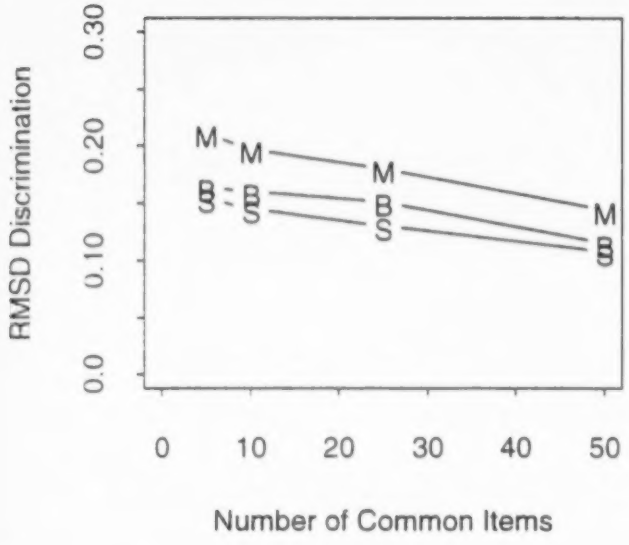


Line of Relationship Between Two Forms

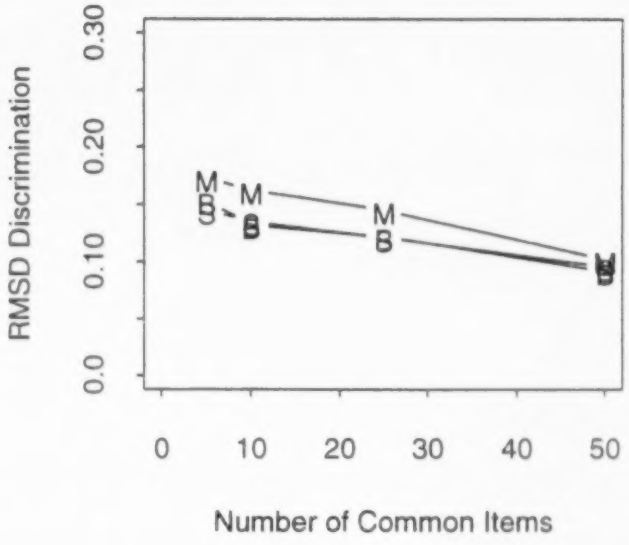


S	Separate Calibration
B	Concurrent BILOG
M	Concurrent MULTILOG

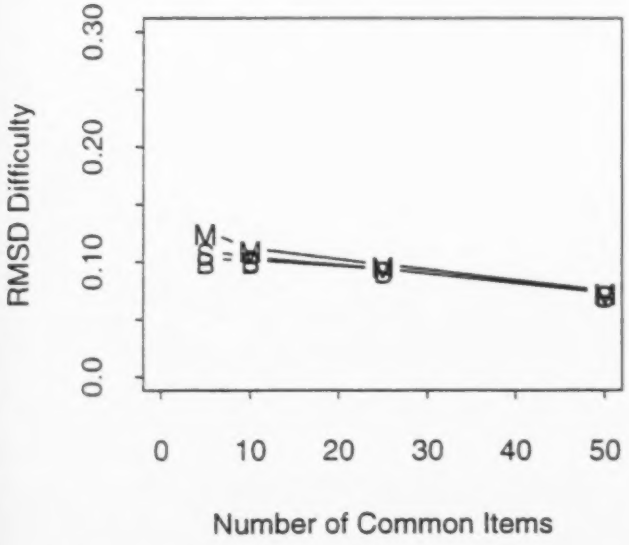
N(0,1) Target Group



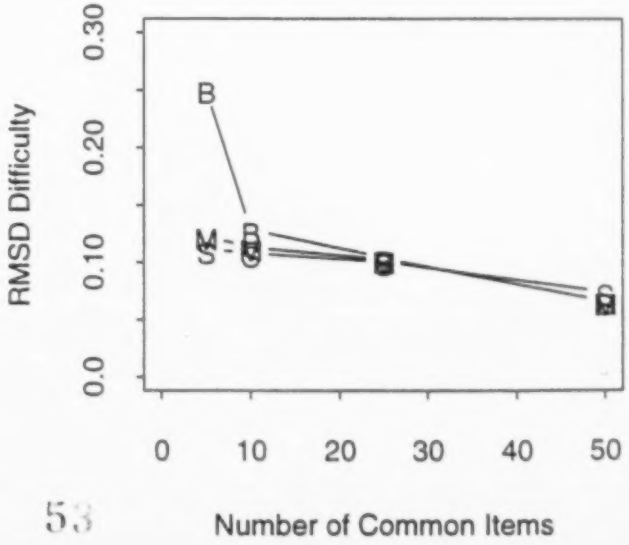
N(1,1) Target Group



N(0,1) Target Group

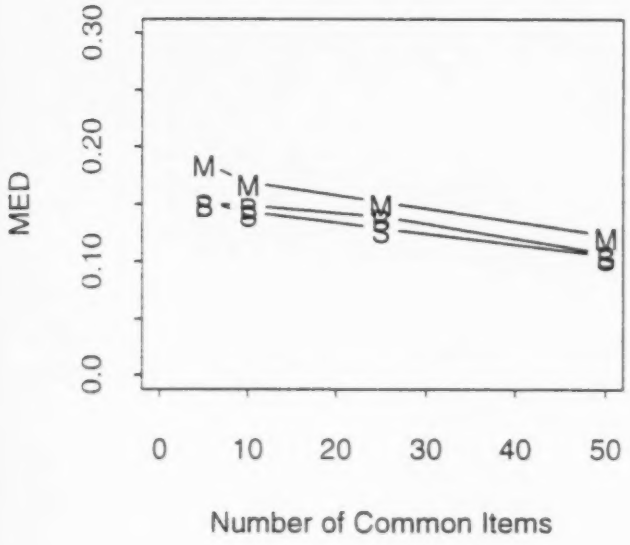


N(1,1) Target Group



S	Separate Calibration
B	Concurrent BILOG
M	Concurrent MULTILOG

N(0,1) Target Group



N(1,1) Target Group